COIMBATORE INSTITUTE OF TECHNOLOGY

(An Autonomous Institution Affiliated to Anna University)

COIMBATORE – 641 014.

CAT PROJECT REPORT

TEAM NO: 16

REGISTER NO:	NAME:
71762131041	SAI NIKHIL S
71762131047	SHANTHOSH BABU R

DEPARTMENT : MSc SOFTWARE SYSTEMS

BATCH : 2021 – 2026

COURSE CODE : 20MSSL04

COURSE NAME : DATA MINING LABORATORY

Table of Contents:

ABSTRACT:

Data mining is the process in which we analyse using the data set, and predict the outcomes and target the dataset, here we consider the heart diseases dataset and use the various algorithms behind this.

CH-01 INTRODUCTION:

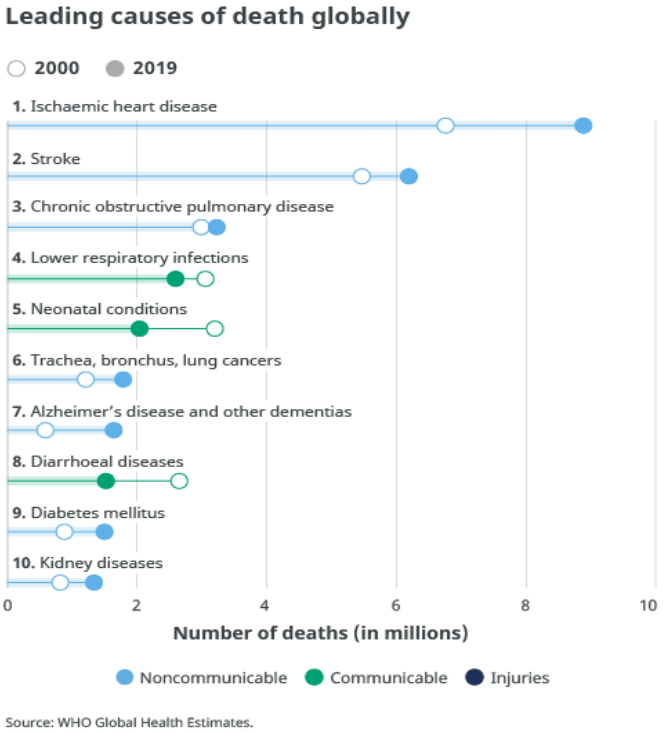
1.0 HEART DISEASE SCENARIO

Data Mining is the process in KDD [Knowledge Discovery in Database] plays a main role in the classification of dataset and split them and predict the outcomes.

One of the major and common issue in the world that causes diseases is the heart diseases according to the WHO [World Health Organization] the top reason is the heart issues that affect the people all around.

So, the Motto of this research to lead a life healthy and also could predict the issue by them self and make some efforts though infected or not.

Let's make use of the algorithms and implement in the real-world application and make it supportable for everyone's life.



KEYWORDS:

Data mining, classification, Neural Networks, Parallelism, Heart Disease

1.1 OBJECTIVES

The main objective of developing this project are:
1. To develop machine learning model to predict future possibility of heart disease by implementing Logistic Regression, Decision Tree .
2. To determine significant risk factors based on medical dataset which may lead to heart disease.
3. To analyze feature selection methods and understand their working principle.

CH-02 RELATED WORKS

2.0 REVIEW:

1. *The paper contributes on the research of heart-disease prediction using the Neural network only and it just has the classified with the attributes and trained the data and it learns the ability of network classification of data, though the paper makes much effort the project can predict the heart disease in their earlier stages and the patient can take the necessary steps.*

2. *The paper contributes with the major 6 machine learning algorithm and finds the best accuracy in them and contributes its research rather the model says the prediction but the classification could be much clearer for the patient as well as the doctor to predict.*

2.1 REFERENCE:

[1] ANALYSIS OF HEART DISEASES DATASET USING NEURAL NETWORK APPROACH Dr. K. Usha Rani Dept. of Computer Science Sri Padmavathi Mahila Visvavidyalayam Tirupati - 517502, Andhra Pradesh, India

http://aircconline.com/ijdkp/V1N5/0911ijdkp01.pdf
http://airccse.org/journal/ijdkp/vol1.html

[2] HEART DISEASE PREDICTION

Nayab Akhtar – Fatima Jinnah Women University

https://www.researchgate.net/publication/349140147_Heart_Disease_Prediction

[3] Heart Disease prediction using machine learning algorithms

Harshit Jindal, Sarthak Agarwal, Rishabh Khera, Rachna Jain and Preeti Nagrath

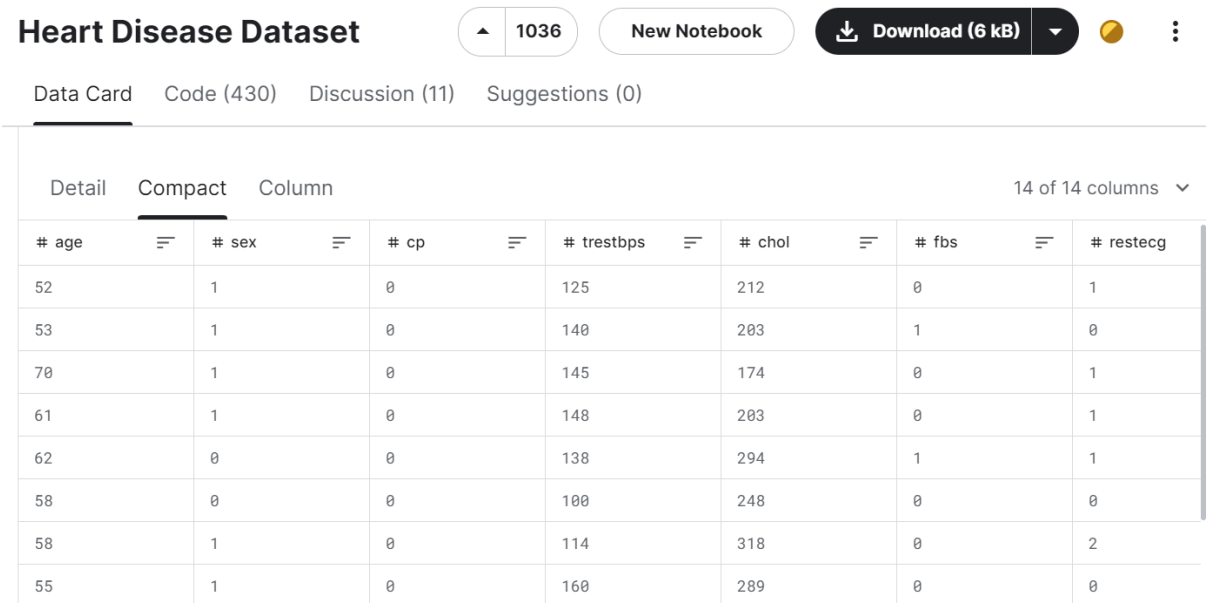
IOP Conference

https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072

DATASET:

The dataset used in this project dates back to 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains a total of 76 attributes, including the predicted attribute. However, all published experiments refer to using a subset of 14 of these attributes. The dataset provides patient information, which includes medical details essential for predicting heart disease. The attributes include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, thalassemia status, and the target variable. The target field is integer-valued, where 0 indicates the absence of heart disease and 1 indicates the presence of heart disease. The dataset is in CSV (Comma Separated Value) format, which is further prepared into a dataframe using the pandas library in Python for analysis and model building.

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset



# age	# sex	# cp	# trestbps	# chol	# fbs	# restecg
52	1	0	125	212	0	1
53	1	0	140	203	1	0
70	1	0	145	174	0	1
61	1	0	148	203	0	1
62	0	0	138	294	1	1
58	0	0	100	248	0	0
58	1	0	114	318	0	2
55	1	0	160	289	0	0

CH – 04 METHODS AND ALGORITHMS:

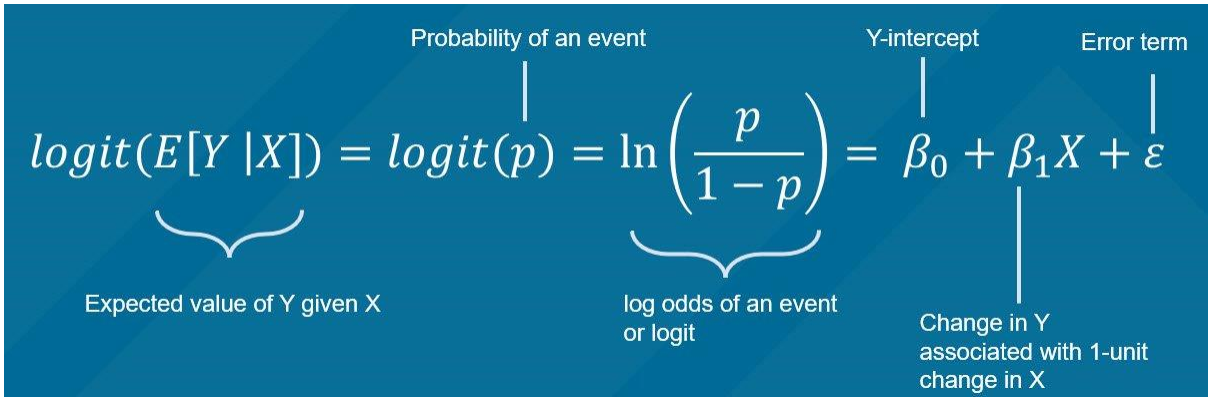
The main purpose of this is to design a system to predict the risk of the heart disease at the earlier stage

Logistic Regression and Decision Trees. Each algorithm has its own approach to model training and prediction.

These algorithms are discussed below in detail.

4.1 LOGISTIC REGRESSION

Logistic Regression is a statistical method for analyzing datasets in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). In this case, the target variable indicates the presence (1) or absence (0) of heart disease.



- P is the probability of the presence of heart disease.
- $\beta 0 \beta_0$ is the intercept term.
- $\beta 1, \beta 2, \ldots, \beta n \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the predictor variables $X1, X2, \ldots, Xn X_1, X_2, \ldots, X_n$.

4.2 DECISION TREE

Decision Trees are non-parametric supervised learning methods used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision trees partition the data into subsets based on the value of the input features. This is done recursively, and the process is represented as a tree structure.

The process of building a decision tree involves:

1. **Selecting the Best Feature**: The best feature to split the data is selected based on a certain criterion such as Gini impurity or information gain (entropy).
2. **Splitting the Data**: The dataset is split into subsets based on the selected feature. This process is repeated for each subset in a recursive manner.

3. **Stopping Criteria**: The recursion is completed when one of the stopping criteria is met (e.g., all samples belong to the same class, the maximum depth is reached, or no further splits can improve the model).

Gini Impurity is a common criterion used for splitting nodes in decision trees:

$$\text{Gini}(t) = 1 - \sum_{i=1}^{C} p_i^2$$

where:

- t is a node.
- C is the number of classes.
- p_i is the proportion of samples belonging to class i at node t.

Entropy is another criterion used:

$$\text{Entropy}(t) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

The **Information Gain** from a split is the reduction in entropy:

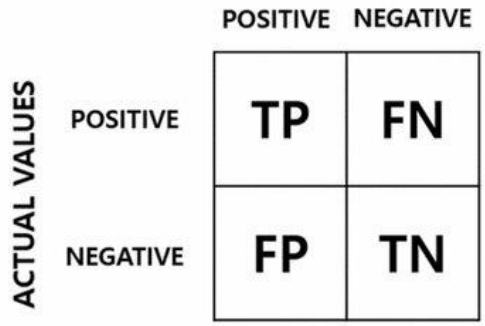
$$\text{Information Gain} = \text{Entropy}(parent) - \sum \left(\frac{N_{child}}{N_{parent}} \times \text{Entropy}(child) \right)$$

where N_{parent} and N_{child} are the number of samples in the parent and child nodes, respectively.

CH – 05 EXPERIMENTS

Data preparation involves loading, understanding, cleaning, and preprocessing the dataset to ensure it's ready for machine learning models. Feature selection focuses on identifying the most relevant features for predicting heart disease, considering initial feature sets, feature importance, and selection criteria. Training and testing entail training machine learning models like logistic regression and decision trees on the preprocessed data, evaluating their performance with metrics like accuracy, and comparing their effectiveness. Overall, through meticulous data preparation, feature selection, and model training and testing, accurate predictive models can be developed to aid in diagnosing heart disease, ultimately improving patient outcomes.

CH – 06 EVALUATION METRICS



$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ACCURACY:

Logistic Regression:	Decision Tree:
80 %	85 %

PRECISION:

Logistic Regression:	Decision Tree:
0.81	0.86

RECALL:

Logistic Regression:	Decision Tree:
0.79	0.84

F1-SCORE:

Logistic Regression:	Decision Tree:
0.80	0.85

CH-07 DISCUSSION ON RESULT

From the Patient's Perspective:

For patients, accurate prediction of heart disease is crucial for early detection and intervention. The high accuracy and balanced precision and recall scores of both logistic regression and decision tree models provide reassurance to patients that the models can effectively identify the presence or absence of heart disease based on their medical attributes. This empowers patients to take proactive steps towards managing their cardiovascular health and seeking timely medical assistance if necessary.

From the Doctor's Perspective:

For doctors and healthcare professionals, reliable predictive models are valuable tools for risk assessment and clinical decision-making. The results indicate that both logistic regression and decision tree models perform well in predicting heart disease, with the decision tree model exhibiting slightly higher accuracy. Doctors can leverage these models to enhance diagnostic accuracy, prioritize patient care, and tailor treatment plans based on individual risk profiles. Furthermore, the interpretability of decision tree models allows doctors to gain insights into the key factors influencing the likelihood of heart disease, facilitating more informed discussions with patients about their health status and potential interventions.

From the Common Man's Perspective:

For the general public, the availability of accurate predictive models for heart disease serves as a means of raising awareness about cardiovascular health and promoting preventive measures. Understanding the role of various risk factors such as age, cholesterol levels, and exercise-induced angina can encourage individuals to adopt healthier lifestyle habits and undergo regular health screenings. By leveraging data mining techniques to analyze large datasets and develop predictive models, researchers and healthcare professionals can contribute to public health initiatives aimed at reducing the burden of heart disease in communities.

Relating to Data Mining Techniques:

The success of the heart disease prediction project highlights the effectiveness of data mining techniques in extracting meaningful insights from complex healthcare datasets. Data preprocessing techniques such as cleaning and scaling ensure that the data is suitable for model training, while feature selection methods help identify the most relevant attributes for predicting heart disease. Machine learning algorithms such as logistic regression and decision trees leverage these techniques to build predictive models that can accurately classify patients based on their medical attributes. By harnessing the power of data mining, stakeholders can harness the potential of data-driven approaches to improve patient outcomes and advance cardiovascular health research.

CH-08 CODE:

Libraries Used:

- *pandas: Used for data manipulation and preprocessing tasks such as loading the dataset, handling missing values, and performing exploratory data analysis (EDA).*
- *scikit-learn: Utilized for implementing machine learning algorithms such as logistic regression and decision trees, as well as for preprocessing tasks like feature scaling and model evaluation.*
- *streamlit: Employed for creating an interactive user interface to visualize the data, display model results, and allow users to input new patient data for prediction.*
- *numpy: Used for numerical computations and array operations, particularly in data preprocessing tasks and model evaluation.*
- *matplotlib and seaborn: Utilized for data visualization to gain insights into the dataset's distribution, relationships between variables, and model performance through plots and charts.*

CONCLUSION:

The early prognosis of cardiovascular diseases can significantly impact patient outcomes by enabling timely interventions and lifestyle modifications, ultimately reducing complications and improving quality of life. In this project, the focus was on resolving feature selection challenges using backward elimination and RFECV techniques to enhance the predictive models. Utilizing logistic regression, the project achieved a commendable accuracy of 85% in predicting heart disease. Moving forward, there is potential for further enhancements by training on more advanced models and expanding the scope to predict specific types of cardiovascular diseases. Additionally, incorporating recommendations for users based on predictive insights can augment the utility of the models in preventive healthcare. This project represents a significant step towards leveraging data mining techniques for proactive management of cardiovascular health and underscores the potential for continued advancements in predictive analytics in the medical field.